ECE 356 Lab 4

Jonathan Chung – jt2chung – 20714889  
Naman Jain – n27jain – 20733273

# Data Set Size Computation

## Task A: Hall of Fame Nomination

Since the classifier is required to identify players who have been nominated, the desired input data set is the set of all players in the database regardless of whether they have been nominated. To find the size of the considered data set, the following SQL query is used:

SELECT COUNT(DISTINCT playerID) "num\_instances" FROM Master;

This query returns a result of 19105. To validate the results of the classifier, the number of players who have been nominated is also computed. This is done with the following SQL query:

SELECT COUNT(DISTINCT playerID) "num\_nominated" FROM HallOfFame;

This query returns a result of 1260. From these numbers, less than 7% of the players in the data set have been nominated for the hall of fame. These queries yield the true sizes of the desired data sets. However, because of various data cleansing issues, these numbers are different from the size of the data set which is used by the classifier. Queries which are more representative of the collected data are:

SELECT COUNT(DISTINCT playerID) "num\_instances" FROM Appearances;

SELECT COUNT(DISTINCT playerID) "num\_nominated"  
FROM HallOfFame JOIN Appearances USING(playerID);

These queries return results of 18912 and 1193, respectively.

## Task B: Hall of Fame Induction

For determining whether players will be inducted, only players who have been nominated will be considered. Since all players who have been nominated are in the HallOfFame table, the size of the data set is given by the following SQL query:

SELECT COUNT(DISTINCT playerID) "num\_nominated" FROM HallOfFame;

This is the same as the second query presented in part A of the query, and as such, produces the same output, 1260. To determine the number of players in this set that have been inducted, the following SQL query is used:

SELECT COUNT(DISTINCT playerID) "num\_inducted"  
FROM HallOfFame WHERE inducted = "Y";

This query returns a result of 317. From these numbers, approximately 25% of the players who have been nominated are then inducted. Similarly to Task A, when accounting for data cleansing issues, these queries become:

SELECT COUNT(DISTINCT playerID) "num\_nominated"  
FROM HallOfFame JOIN Appearances USING(playerID);

SELECT COUNT(DISTINCT playerID) "num\_inducted"  
FROM HallOfFame JOIN Appearances USING(playerID) WHERE inducted = "Y";

These queries then return 1193 and 251, respectively.

# Data Cleansing Issues

A variety of data cleansing issues were encountered in the dataset. Firstly, not every player is present in every table. For example, there is a player (with playerID = “drewj.01”) in the HallOfFame table who does not appear in the Master table. This makes it difficult to select a data set containing all the players regardless of whether they have been nominated. Furthermore, for a player to be classified by the classifier, there must be some data on their performance, such as the number of games they have played in. This data is provided by the Appearances table. However, there are 67 playerIDs which appear in the HallOfFame table but not the Appearances table. To deal with this issue, only players who have an entry in the Appearances table are considered, filtering out players who do not have statistics in the database. Notably, this reduces the size of the data set. The Appearances table contains 18912 different players, while the Master table contains 19105 players.

A similar issue is the fact that not every player occurs in both the Batting and Pitching tables. To resolve this, left outer joins are used to combine these tables with the base data sets. However, this results in null values in columns from these tables. To deal with this, null values are mapped to a value 0 using IFNULL(attribute, 0). A value of 0 is used because it is assumed that players do not appear in these tables if they have never batted or pitched.

Another issue which was encountered when generating the verification data for Task B is that a single playerID can appear in the HallOfFame table multiple times, with different values of the inducted attribute. This is the result of the fact that players can be nominated for the HallOfFame multiple times, and although they do not get inducted initially, they are inducted upon a subsequent nomination. Thus, to correctly generate the data for whether a player has been inducted, the following SQL snippet is used as a subquery:

SELECT playerID, inducted   
FROM HallOfFame  
WHERE playerID NOT IN (  
 SELECT playerID FROM HallOfFame WHERE inducted = "Y"  
) UNION (  
 SELECT playerID, inducted FROM HallOfFame WHERE inducted = "Y"  
)

# Feature Selection

TODO

# Data Sampling

TODO

# Validation Results

TODO

# Hyper Parameter Tuning

## Task A: Hall of Fame Nomination

TODO

## Task B: Hall of Fame Induction

TODO

# Decision Trees

## Task A: Hall of Fame Nomination

TODO

## Task B: Hall of Fame Induction

TODO